

R

ENDEZ-VOUS

P.80 Logique & calcul
P.86 Art & science
P.88 Idées de physique
P.92 Chroniques de l'évolution
P.96 Science & gastronomie
P.98 À picorer

INTELLIGENCES ARTIFICIELLES : UN APPRENTISSAGE PAS SI PROFOND !

Les systèmes de reconnaissance automatique ont d'étonnantes faiblesses. Exploiter ces failles permet de s'amuser, mais aussi d'améliorer les procédures d'apprentissage... ou de concevoir de nouvelles attaques malveillantes.

L'AUTEUR



JEAN-PAUL DELAHAYE
professeur émérite
à l'université de Lille
et chercheur au Centre
de recherche en
informatique, signal
et automatique de Lille
(Cristal)



Jean-Paul Delahaye a récemment publié : **Les Mathématiciens se plient au jeu**, une sélection de ses chroniques parues dans *Pour la Science* (Belin, 2017).

L'apprentissage automatique à l'aide de réseaux neuronaux «profonds» est à la mode. Comme souvent quand il s'agit d'intelligence artificielle, il y a ceux qui perçoivent raisonnablement les capacités des nouvelles idées et il y a ceux qui imaginent bien plus. Cet enthousiasme excessif s'est produit avec les premiers succès des méthodes d'apprentissage automatique dans la décennie 1950; cela a recommencé avec les systèmes experts dans les années 1980; puis quelque temps après avec les réseaux de neurones; et maintenant, c'est une forme de ces réseaux qui fait croire que nous sommes sur le point de mettre dans nos ordinateurs une intelligence générale susceptible de nous surpasser.

Le présent article n'a pas pour objectif de dénigrer une remarquable technique qui a récemment triomphé au jeu de go, qui aide à concevoir des véhicules autonomes, qui améliore la traduction automatique et qu'on maîtrise de mieux en mieux; les livres d'Aurélien Géron (*voir la bibliographie*) vous initieront à cette science nouvelle. Nous cherchons ici à remettre les pieds sur terre à ceux qui l'imaginent comme une panacée informatique.

RÉSEAUX NEUROMIMÉTIQUES

Les neurones formels proposés en 1959 sont des modèles informatiques et simplifiés des neurones de notre cerveau. On conçoit des programmes qui en simulent un grand nombre en les regroupant en couches successives, comme les neurones du cerveau humain. Les neurones

formels de la couche d'entrée du réseau reçoivent des informations, par exemple sous la forme d'images décomposées en pixels: chaque pixel est une entrée. En fonction de leurs paramètres internes, les neurones de cette première couche envoient des signaux aux neurones de la seconde couche, qui eux-mêmes, en fonction de leurs paramètres internes, envoient des signaux à la troisième couche, etc.

Les influx de sortie, c'est-à-dire de la dernière couche, désignent les réponses possibles. Ce sont par exemple des lettres A, B, C ...: si les images données en entrée sont des lettres manuscrites et imprécises, le but recherché est que le réseau, une fois instruit, sache correctement reconnaître les lettres qui lui sont proposées. La phase d'apprentissage consiste à ajuster les paramètres internes de chaque neurone pour que le réseau réponde correctement. Souvent, plusieurs milliers de paramètres sont à calculer. Tout l'art de la programmation des réseaux de neurones consiste, à l'aide d'un grand nombre d'images avec leurs classifications (par exemple des formes manuscrites dont on indique au réseau à quelles lettres elles correspondent), à faire converger les jeux de paramètres des neurones vers un état à peu près stabilisé et qui sera tel que lorsqu'on proposera des images, par exemple d'autres lettres manuscrites non utilisées, elles soient correctement identifiées.

On parle d'apprentissage supervisé: lors d'une phase initiale, on indique au réseau les réponses voulues; il apprend et, si tout s'est bien passé, il devient alors capable de trouver, seul, les bonnes réponses à des entrées

PERFORMANCES : L'EXEMPLE DU SYSTÈME CAPTIONBOT

1

L'identification automatique d'objets sur une photo a fait récemment de grands progrès et les systèmes d'aujourd'hui sont capables, avec un certain succès, de reconnaître le contenu d'images délicates. Vous vous en rendez compte vous-même en testant le système en ligne CaptionBot conçu par Microsoft (<https://www.captionbot.ai>).

Le système combine deux réseaux neuronaux. Le premier s'occupe de reconnaissance d'images,

le second du traitement du langage naturel pour composer les réponses. En prenant en compte un grand nombre d'images accompagnées de leur description, le logiciel apprend à associer les caractéristiques de l'image à des descriptions écrites de ce qu'elles montrent. Il tente alors de faire de même avec de nouvelles images. Le système continue de progresser. En février 2018, on obtenait par exemple les résultats ci-dessous (les réponses ont été traduites).



BON
Je pense que c'est un arbre avec en fond un coucher de soleil.



BON
Je pense que c'est une personne tenant une guitare.



BON
Je pense que c'est une femme assise sur un canapé.



PAS SI MAL
Je pense que c'est un homme sur une plage.



PAS SI MAL
Je n'en suis pas très sûr, mais je pense que c'est un ours brun assis sur une table.



RATÉ
Je pense que c'est un grand étalage de maïs.



RATÉ
Je pense que c'est une paire de chaussures bleues.



RATÉ
Je n'en suis pas très sûr, mais je pense que c'est un avion dans le ciel.

nouvelles. Un autre type de méthodes pour instruire un réseau de neurones formels se fonde sur l'idée de le récompenser quand il propose une bonne réponse et de le pénaliser sinon; ce sont les méthodes d'«apprentissage par renforcement».

La puissance de calcul disponible aujourd'hui, les bases de données volumineuses collectées pour les exemples utilisés en phase d'apprentissage, ainsi que la multiplication des couches internes d'un réseau (on en utilise parfois plusieurs dizaines) ont conduit récemment à de spectaculaires réussites, comme la victoire des machines sur les humains au jeu de go.

Les chercheurs du domaine étaient persuadés qu'un réseau ayant bien appris était nécessairement robuste et doté d'une propriété de continuité: des traits dessinant de manière vague et approchée une lettre seront reconnus correctement, et une image proche d'une image reconnue sera elle aussi reconnue.

Or ce n'est pas le cas: les chercheurs ont mis au point des images pièges, conçues pour qu'un réseau ayant appris à reconnaître une certaine catégorie d'images propose une réponse incongrue et fautive quand on lui soumet une image proche, qu'un humain classe aisément.

>